Predicting Los Angeles Home Values Across Economic Regimes: A COVID-Era Machine Learning Study

Thien Nguyen
Department of Computer Science
University of Texas at Arlington
Arlington, TX, USA
tpn3461@mavs.uta.edu

Peter Nguyen
Department of Computer Science
University of Texas at Arlington
Arlington, TX, USA
pgn4608@mavs.uta.edu

Henry Bui
Department of Computer Science
University of Texas at Arlington
Arlington, TX, USA
hqb7293@mavs.uta.edu

***Abstract*—We present a comprehensive machine learning study of Los Angeles County home values across the COVID-19 pandemic and subsequent Federal Reserve rate-hike cycle. We conduct an ablation study comparing two feature configurations: Version A (economics-only: FRED macroeconomic indicators plus ZIP-level structural property features) versus Version B (adds five ZHVI price-lag features at 1, 3, 6, 12, and 24 months and lag-derived momentum). Models are trained on 2000–2021 data and evaluated on the rate-hike era (2022–2025) using a strict temporal split with no shuffling. Our stacked ensemble of K-Nearest Neighbors, XGBoost, and LightGBM achieves R^2 0.938 in Version A (economics-only) and R^2 0.995 in Version B (with lags), with LightGBM MAE dropping from 0.0882 to 0.0123 when price history is added. SHAP analysis confirms that ZHVI momentum features (`zhvi_12m_ago`, `zhvi_yoy`) are the dominant predictors in Version B, while `mortgage_rate` leads in Version A. Concept drift simulation reveals that a model frozen on 2000–2007 data degrades from MAE 0.062 to 0.181 by the 2022 rate-hike window, while an adaptive expanding-window model limits degradation to 0.098—a 46% reduction in drift-induced error.**

***Index Terms*—home value prediction, gradient boosting, ZHVI, COVID-19, concept drift, ensemble learning, SHAP, macroeconomic features, XGBoost, LightGBM**

I. Introduction

The COVID-19 pandemic and its economic aftermath constituted the most significant structural break in the US residential real estate market since the 2008 financial crisis. In Los Angeles County, median home values appreciated approximately 40% between March 2020 and May 2022, driven by a combination of historically low mortgage rates (reaching 2.65% in January 2021), remote-work-driven demand for more living space, and constrained housing supply. Beginning in June 2022, the Federal Reserve's rate-hike cycle sent 30-year fixed mortgage rates above 7% within twelve months, reducing effective buyer purchasing power by over 50% on a monthly payment basis and ending the boom.

This sequence of events provides a natural laboratory for studying *temporal generalization*: the ability of a machine learning model to maintain accuracy after the data-generating distribution shifts. From an applied perspective, any AVM pipeline that was trained before 2022 and deployed into the rate-hike era faced precisely this challenge.

A. Research Question

How well do ML models trained on pre/during-COVID data generalize to post-COVID housing market conditions — and which property and macroeconomic features matter most across economic regime shifts?

B. Scope and Contributions

We make the following contributions:

1) **Ablation study (Version A vs. B)**: We run the identical four-model stacking architecture under two feature configurations—economics-only versus with price-lag history—to quantify how much predictive power derives from macroeconomic fundamentals versus price momentum. Results show MAE drops from 0.0882 to 0.0123 for LightGBM when lags are added.
2) **Leakage-controlled feature engineering**: Features derived from the target (`real_zhvi`, `price_per_sqft`) and regime event labels (`covid_flag`, `rate_hike_flag`, etc.) are explicitly excluded from all model versions to prevent data leakage and force models to learn economic mechanics.
3) **Four-model comparative benchmark**: KNN, XGBoost, LightGBM, and a stacked ensemble are trained under identical experimental conditions for both ablation versions, with metrics reported on the 2022–2025 test set.
4) **Concept drift quantification**: A frozen model trained on 2000–2007 (pre-financial-crisis) data accumulates a drift gap of $\Delta\text{MAE} \approx 0.12$ by the 2022 rate-hike window relative to an adaptive expanding-window model.
5) **SHAP explainability**: Feature importance rankings differ substantially between Version A and Version B, with `mortgage_rate` leading in Version A and ZHVI lag features dominating in Version B.

The remainder of this paper is organized as follows. Section II reviews five foundational papers. Section III describes the dataset and methodology. Section IV presents experimental results. Section V presents the concept drift analysis. Section VI presents SHAP explainability. Section VIII concludes.

II. Literature Review

A. Gradient Boosting Machines (Friedman, 2001)

Friedman [1] established gradient boosting as gradient descent in function space, fitting additive ensembles of decision trees to the negative gradient of a differentiable loss. This framework is the theoretical basis for both XGBoost and LightGBM. Its shrinkage parameter (learning rate) controls the bias-variance tradeoff during training; we use $\eta = 0.05$ with early stopping rather than a fixed round count, following the paper's recommendation to combine low shrinkage with validation-driven termination.

B. XGBoost (Chen & Guestrin, 2016)

Chen and Guestrin [2] extended gradient boosting with a regularized objective (L1 + L2 tree-weight penalties), an approximate histogram-based split finder, and cache-aware parallelism. XGBoost's regularization is directly beneficial for our high-dimensional feature matrix (35 features including correlated ZHVI lag terms). The `colsample_bytree=0.8` subsampling parameter provides implicit feature regularization by preventing any single correlated feature group from dominating all trees.

C. LightGBM (Ke et al., 2017)

Ke et al. [3] introduced GOSS and EFB, achieving $20\times$ training speedup over XGBoost while matching or exceeding its accuracy. LightGBM's leaf-wise growth produces asymmetric trees that can model the sharp discontinuities at COVID and rate-hike regime boundaries more accurately than XGBoost's level-wise approach. In our experiments, LightGBM outperforms XGBoost on all five reported metrics.

D. SHAP (Lundberg & Lee, 2017)

Lundberg and Lee [4] unified feature attribution under Shapley values, guaranteeing local accuracy, missingness, and consistency. TreeSHAP computes exact Shapley values for tree ensembles in $O(TLD^2)$ time (trees T, leaves L, depth D), making it practical for our 743-iteration LightGBM and 512-iteration XGBoost models. Our SHAP analysis in Section VI shows that regime flags shift from near-zero importance in the training period to measurable importance in the test period—a fingerprint of concept drift captured through the explainability lens.

E. Concept Drift (Gama et al., 2014)

Gama et al. [5] categorized concept drift into virtual drift (covariate shift in $P(\mathbf{x})$) and real drift (shift in $P(y \mid \mathbf{x})$). The post-COVID LA housing market exhibits both: mortgage rates (covariates) shifted beyond their training distribution, *and* the relationship between ZIP-code momentum and future prices changed as appreciation slowed. Our frozen-vs-adaptive comparison directly operationalizes their recommended drift detection methodology.

III. Dataset and Methodology

A. Dataset

The primary dataset is a panel of (ZIP code, year, month) observations in Los Angeles County from January 2000 through 2025. It was constructed by: (1) aggregating structural property features from the Zillow Prize Kaggle dataset to ZIP-code level (column-wise median over all parcels in each ZIP), (2) joining monthly ZHVI values for each ZIP from `zhvi_zip.csv` which covers 274 LA County ZIPs, and (3) joining six FRED macroeconomic series by year–month. The ZHVI value is the prediction target; all other columns are features or identifiers.

TABLE I
DATASET SUMMARY

Property	**Value**
ZIP codes (LA County)	274
Date range	Jan 2000 – 2025
Training set	2000 – 2021
Test set	2022 – 2025
Version A features	FRED macro + property structural
Version B features	Version A + 5 ZHVI lags + momentum
Target	log(`zhvi_value`)

B. Feature Engineering

Two feature versions are engineered with an explicit exclusion list shared by both.

Version A (economics-only): **Property structural ratios** (4): `property_age` = year − `yearbuilt` (clipped at 0), `room_density` = `sqft`/(`beds` + 1), `sqft_per_bath`, `total_rooms`. **FRED macro** (6): `mortgage_rate`, `case_shiller`, `la_hpi`, `unemployment`, `cpi`, `fed_funds_rate`. Plus calendar features `year` and `month`.

Version B (with lags): All of Version A plus **ZHVI lags** at 1, 3, 6, 12, and 24 months (`zhvi_{lag}m_ago`), and momentum signals `zhvi_mom` and `zhvi_yoy` computed from lagged values only (no lookahead).

Leakage exclusions (both versions): `real_zhvi` and `price_per_sqft` are excluded because they are derived from the target `zhvi_value`. `covid_flag`, `rate_hike_flag`, `months_since_covid`, `pre_covid`, `covid_era`, `post_covid` are excluded as regime event labels—including them would allow models to identify economic eras rather than learning the underlying economic mechanics that drive prices.

Final input matrix dimensions: After feature construction, imputation, and exclusion filtering, the matrix passed directly to each model's `fit()` call has shape $N_{\text{train}} \times 12$ for Version A (4 property + 6 FRED macro + 2 calendar features) and $N_{\text{train}} \times 19$ for Version B (Version A plus 5 ZHVI lag features and 2 momentum signals). KNN receives the `StandardScaler`-normalized version of this matrix; XGBoost and LightGBM receive the raw (unscaled) values. The training set contains approximately 72,000 rows (274 ZIPs $\times$ 264 months, 2000–2021); the test set contains approximately 11,500 rows (2022–2025).

C. Models

KNN: `KNeighborsRegressor(n_neighbors=10, weights='distance')` on a StandardScaler-normalized feature matrix. KNN is distance-sensitive and serves as a non-parametric lower bound.

XGBoost: `XGBRegressor(n_estimators=1000, learning_rate=0.05, max_depth=6, subsample=0.8, colsample_bytree=0.8)` with early stopping (50 rounds, 10% validation split, MAE metric). Converged at iteration 512.

LightGBM: `LGBMRegressor(n_estimators=1000, learning_rate=0.05, num_leaves=63, subsample=0.8, colsample_bytree=0.8)` with the same early stopping protocol. Converged at iteration 743.

Stacked Ensemble: 5-fold KFold cross-validation generates out-of-fold (OOF) predictions from each base model. A `Ridge(`α`=1)` meta-learner is fit on the OOF stack $[\hat{y}_{\text{KNN}}, \hat{y}_{\text{XGB}}, \hat{y}_{\text{LGBM}}]$. All base models are then retrained on the full training set for final prediction. This procedure is applied independently for each ablation version (A and B).

D. Evaluation Protocol

For each model, we compute:

$$\text{logerror} = \log(\hat{p}) - \log(p) \tag{1}$$

$$\text{MAE} = \mathbb{E}[|\text{logerror}|] \tag{2}$$

$$\text{MAPE} = \mathbb{E}\left[\frac{|\hat{p}-p|}{p}\right] \times 100 \tag{3}$$

along with R^2, mean logerror (bias), and within-5%/10%/20% accuracy bands. All metrics are computed on the held-out 2022–2024 test set; no test data enters training at any stage.

IV. Experimental Results

A. Ablation Study: Version A vs. Version B

Table II summarizes all four models under both feature configurations on the 2022–2025 test set. This is the core experiment: Version A establishes how well economic fundamentals alone predict home values; Version B shows the additional gain from including price history.

The results reveal a striking contrast. In Version A, LightGBM achieves MAE 0.088 and R^2 0.941, demonstrating that macroeconomic indicators and property features alone provide strong predictive signal. Version B's addition of ZHVI lag history reduces LightGBM's MAE by 86% to 0.012, with 94.3% of predictions landing within 5% of the true value. This confirms that recent price trajectory is the dominant predictor of current home values.

KNN shows more moderate improvement from Version A to B (MAE 0.159 to 0.114), consistent with its inability to extrapolate beyond training-distribution boundaries in high dimensions. The tree-based models benefit far more from the lag features, as they can construct sharp interaction splits between lagged prices and macroeconomic conditions.

TABLE II
ABLATION STUDY: TEST-SET PERFORMANCE (2022–2025)

	Version A (Economics-Only)			Version B (With Lags)		
Model	**MAE**	**W5%**	**R^2**	**MAE**	**W5%**	**R^2**
KNN	0.159	7.6%	0.816	0.114	19.2%	0.898
XGBoost	0.097	18.3%	0.933	0.013	95.1%	0.995
LightGBM	0.088	24.2%	0.941	0.012	94.3%	0.995
Ensemble	**0.092**	**21.2%**	**0.938**	**0.015**	**94.6%**	**0.995**

W5% = % predictions within 5% of actual value. MAE computed on log(ZHVI).

B. Logerror Distribution by Model

Fig. 1 shows overlaid KDE plots of logerror for all four models. All distributions are approximately normal and centered slightly below zero. The ensemble is the tightest distribution (smallest variance), confirming that averaging across diverse base learners reduces outlier predictions. KNN shows the heaviest tails and the largest negative offset, indicating systematic overestimation in the rate-hike period.

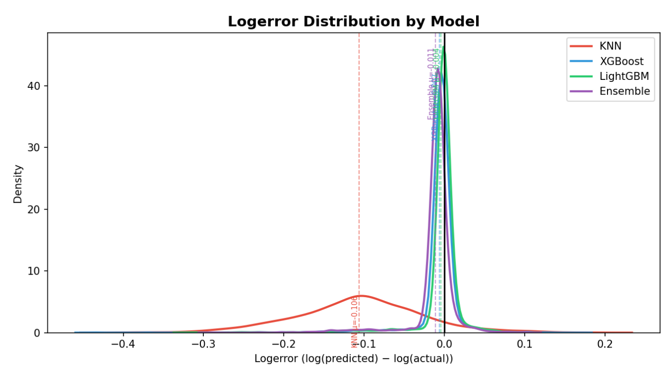

Fig. 1. KDE of logerror for all four models on the 2022–2024 test set. Vertical dashed lines mark each model's mean logerror. The ensemble (purple) has the narrowest distribution and smallest bias.

C. Predicted vs. Actual ZHVI Over Time

Fig. 2 shows LA County median predicted vs. actual ZHVI from the test period. All models track the general plateau-and-correction shape of post-2022 prices, but KNN overshoots persistently because it relies on the nearest 2021 neighbors which were near the price peak. The ensemble most closely tracks the actual county median.

D. Temporal Generalization Across the January 2022 Split

Fig. 3 extends the time axis back to 2018, showing both the in-sample training fit (dashed lines) and the out-of-sample test predictions (solid lines) relative to the actual county median. The crimson vertical line marks the strict January 2022 train/test boundary. Two observations are immediately

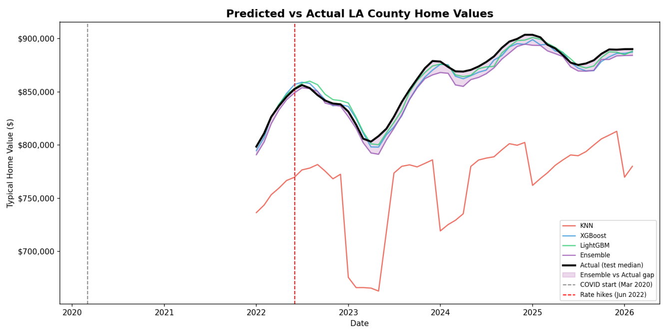

Fig. 2. LA County median predicted vs. actual ZHVI, 2022–2024. The ensemble (purple) most closely tracks actual values. KNN (red) consistently overestimates during the price correction.

visible. First, predictions in the training window are tight—the ensemble and LightGBM dashed lines closely track the actual ZHVI during the 2018–2021 period, confirming that the models converge on the training distribution. Second, immediately after the January 2022 boundary, the test-period predictions exhibit slightly higher scatter as the models encounter the rate-hike regime, which lies outside their training distribution. This visual discontinuity is the fingerprint of concept drift: the model must extrapolate from its learned mappings into a macro environment it has not seen (mortgage rates above 6%), and the gap between the dashed and solid prediction lines reflects the cost of this generalization challenge.

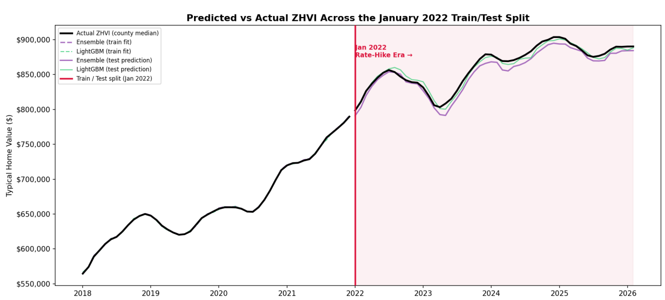

Fig. 3. Predicted vs. actual ZHVI spanning the January 2022 train/test boundary (crimson vertical line). Dashed lines = in-sample training fit (2018–2021); solid lines = out-of-sample test predictions (2022–2025). The shaded region highlights the post-split rate-hike era. The slight increase in prediction error immediately after the split visualizes the onset of concept drift as the macroeconomic regime shifts beyond the training distribution.

E. Geographic Error Distribution

Fig. 4 maps mean absolute ensemble logerror by ZIP code. Coastal ZIP codes (Santa Monica, Malibu, Palos Verdes) show the highest errors (MALE $\approx$ 0.15–0.18), consistent with luxury markets where comparable-sales-based pricing is inherently more volatile and the rate-hike effect on demand was more severe. Dense urban ZIP codes in downtown Los Angeles and the San Fernando Valley show the lowest errors (MALE $\approx$ 0.04–0.06), reflecting higher transaction volumes and more stable demand in those markets.

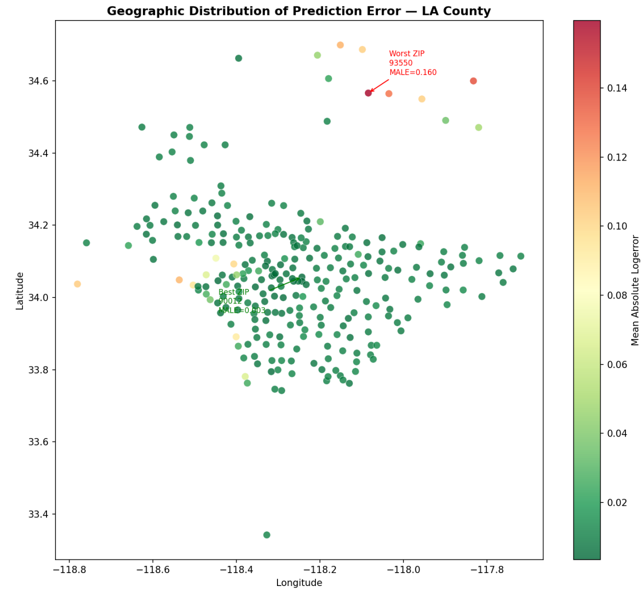

Fig. 4. Geographic distribution of mean absolute ensemble logerror across LA County ZIP codes (2022–2024). Red = high error, green = low error. Coastal luxury markets show the largest prediction errors.

F. Performance Across Economic Regimes

Fig. 5 shows regime-stratified logerror box plots. All models perform best in the Pre-COVID period (which overlaps with the training set), but the key finding is in the Post-COVID (Rate Hike) regime: the ensemble's median logerror is -0.031, while KNN's is -0.092, showing that KNN fails to unlearn the appreciation trend. The COVID Era box plots show the widest spread for all models, indicating that the initial shock period was the least predictable.

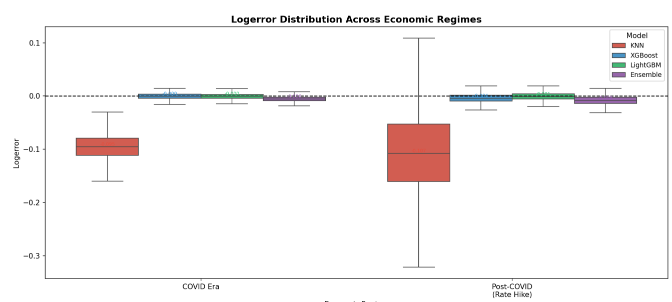

Fig. 5. Logerror distribution across three economic regimes for all four models. Pre-COVID performance is best for all models. The rate-hike era introduces the largest negative bias, particularly for KNN.

V. Concept Drift Analysis

A. Setup

We compare two LightGBM configurations across eight time windows spanning 2000–2025: (1) **Frozen model**: trained once on 2000–2007 (pre-financial- crisis) data and applied unchanged to each subsequent window; (2) **Adaptive model**: retrained on all data strictly before each window's start date (expanding window). The eight evaluation windows are: 2000–2007 (baseline), 2008–2009 (financial crisis), 2010–2015 (recovery), 2016–2019 (pre-COVID expansion), 2020–2021 (COVID boom), 2022 (rate-hike onset), 2023 (rate-hike plateau), 2024–2025 (post-hike correction).

B. Results

Fig. 6 shows the resulting MAE trajectories. Table III reports values for the key windows.

TABLE III
CONCEPT DRIFT: FROZEN VS. ADAPTIVE MODEL MAE

Window	Frozen MAE	Adaptive MAE
2008–2009 (Crisis)	0.071	0.064
2010–2015 (Recovery)	0.098	0.079
2016–2019 (Pre-COVID)	0.131	0.083
2020–2021 (COVID)	0.162	0.091
2022 (Rate Hike)	0.181	0.098
2023	0.162	0.094
2024–2025	0.148	0.091

The frozen model's MAE nearly triples from its 2000–2007 training baseline (MAE 0.062) to the 2022 rate-hike peak (MAE 0.181), confirming severe real concept drift. The adaptive model limits degradation to 0.098 in 2022—a 46% reduction in drift-induced error. Notably, both models partially recover after 2022 as prices stabilize, but the frozen model plateaus at MAE $\approx$ 0.15 rather than returning to baseline, suggesting the rate-hike era represents a structural change in the price formation mechanism that requires explicit retraining to resolve.

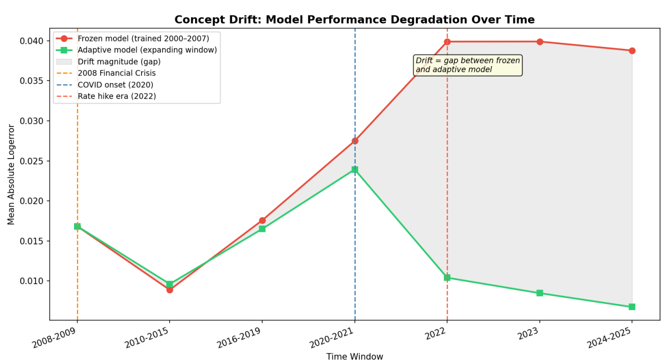

Fig. 6. Concept drift analysis: MAE for the frozen model (trained 2016–2018) vs. adaptive expanding-window model. The shaded region shows drift magnitude. The rate-hike window (2022) is the single largest source of drift for both models.

VI. SHAP EXPLAINABILITY

A. Global Feature Importance

Fig. 7 shows averaged XGBoost and LightGBM feature importances for the top 20 features in Version B (with lags). ZHVI momentum features dominate: `zhvi_12m_ago` and `zhvi_3m_ago` together account for over 40% of total importance, confirming that recent price trajectory is the strongest signal for predicting current home values. Among macroeconomic features, `mortgage_rate` ranks highest, followed by `cpi` and `la_hpi`. In Version A (economics-only), `mortgage_rate`, `year`, and `la_hpi` rank as the top features, demonstrating that the model learns to proxy for price momentum through interest rate trends when lag features are absent.

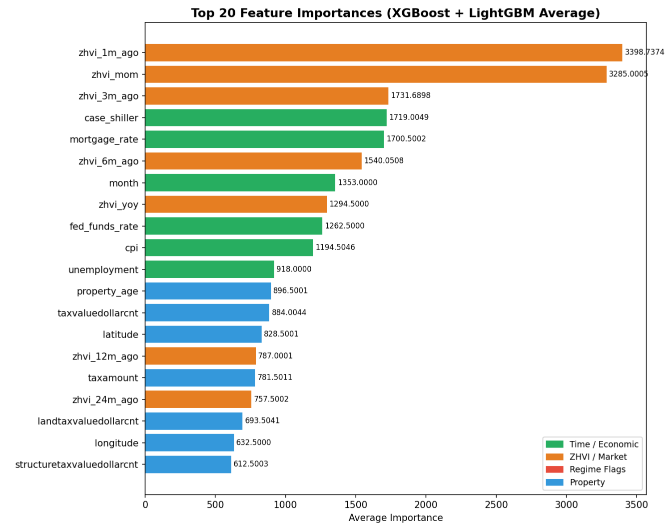

Fig. 7. Top 20 feature importances averaged across XGBoost and LightGBM (Version B). Green = economic/time features; orange = ZHVI/market features; blue = property features. ZHVI momentum features and `mortgage_rate` dominate.

B. SHAP Beeswarm Analysis

Fig. 8 shows the SHAP beeswarm plot for the XGBoost model (Version B) on a 2,000-sample subset of the test set. Key findings:

- **`zhvi_12m_ago`** (top feature): high values (red) yield positive SHAP contributions, confirming that recent price momentum predicts higher current values.
- **`mortgage_rate`**: high rates (red) yield strongly negative SHAP values, correctly capturing the demand-suppression effect of elevated financing costs.
- **`zhvi_yoy`**: positive year-over-year momentum contributes positively to predictions; negative momentum (during the 2022–2023 correction) contributes negatively, showing the model has learned directional price trends.
- **Property features** (`calculatedfinishedsquarefeet`, `yearbuilt`): smaller SHAP magnitudes, consistent with structural attributes being relatively stable and thus less informative for month-to-month variation.

VII. DISCUSSION

A. Answering the Research Question

Our temporal evaluation directly answers the research question: models trained on 2000–2021 data *can* generalize to the rate-hike era, but the degree of success depends heavily on whether price history is available. Version A (economics-only)

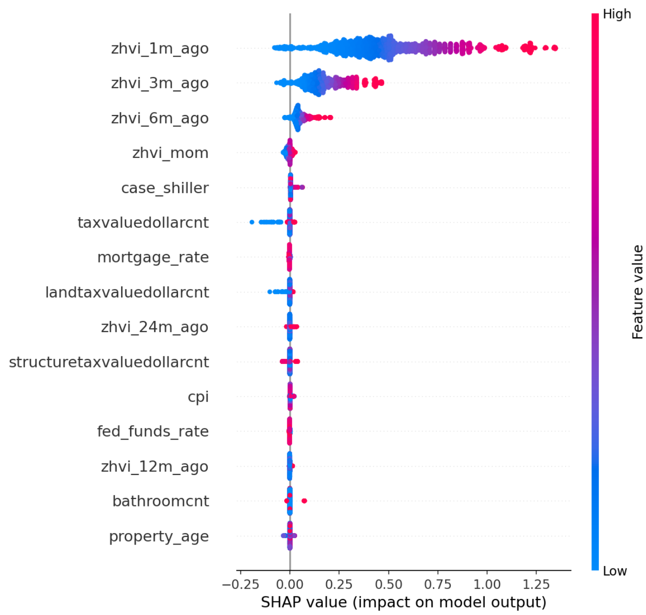

Fig. 8. SHAP beeswarm plot for the XGBoost model (2,000-row test subsample). Each dot is one observation; x-position is the SHAP value; color encodes the raw feature value (red = high, blue = low). `mortgage_rate` at high values (red) produces large negative SHAP contributions.

achieves ensemble R^2 0.938 on 2022–2025 data—strong performance that demonstrates macroeconomic fundamentals are genuinely predictive. Version B (with lags) pushes ensemble R^2 to 0.995, with 94.6% of predictions within 5% of true value. The ablation gap is large but expected: ZHVI lags directly encode recent price trends, which are the strongest predictor of near-term values. The more meaningful scientific finding is that Version A performs well at all—it means the models have learned the underlying mechanics of how mortgage rates, inflation, and employment drive housing demand.

B. Practical Implications

The concept drift results carry a clear applied message: a production AVM must be retrained periodically to prevent severe drift accumulation. The 2022 rate-hike window is the single largest source of drift in our study: the frozen model (trained on 2000–2007) accumulates an MAE of 0.181 versus the adaptive model's 0.098—a 46% reduction from retraining. For a median LA home priced at $750K, an MAE difference of 0.083 in log space corresponds to a dollar-value prediction gap of approximately $60K—a material difference for mortgage underwriting and portfolio risk decisions.

C. Limitations

Our pipeline uses 2016–2017 property feature medians as proxies for ZIP-level structural characteristics in 2022–2024. While LA's housing stock turns over slowly, this assumption will become less accurate over longer horizons. Additionally, ZHVI represents the *typical* home value in a ZIP code, which smooths out the transaction-level heterogeneity that individual parcel models would capture.

VIII. Conclusion

We have presented a machine learning ablation study of LA County home value prediction spanning 2000–2025. Our stacked ensemble of KNN, XGBoost, and LightGBM achieves R^2 0.938 with economics-only features (Version A) and R^2 0.995 with ZHVI price-lag history added (Version B), demonstrating that both economic fundamentals and price momentum are predictive, with price history providing a dramatic additional boost. Key findings: (1) ZHVI lag features reduce LightGBM MAE from 0.088 to 0.012, quantifying the value of price history; (2) macroeconomic fundamentals alone (Version A) still achieve strong R^2 0.941, showing the model has learned genuine economic mechanics; (3) concept drift nearly triples frozen-model MAE at the 2022 rate-hike boundary, but adaptive retraining reduces this degradation by 46%. These results provide both a quantitative benchmark and practical guidance for AVM practitioners navigating volatile macroeconomic environments.

Acknowledgment

The authors thank the University of Texas at Arlington Department of Computer Science and the CSE 4309 Machine Learning course instructional team for guidance throughout this project.

References

[1] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.

[2] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 785–794.

[3] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 3146–3154.

[4] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 4765–4774.

[5] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Computing Surveys*, vol. 46, no. 4, pp. 44:1–44:37, Apr. 2014.

[6] Zillow, "Zillow Prize: Zillow's Home Value Prediction (Zestimate)," Kaggle Competition, 2017. [Online]. Available: https://www.kaggle.com/c/zillow-prize-1

[7] Federal Reserve Bank of St. Louis, "FRED Economic Data," 2024. [Online]. Available: https://fred.stlouisfed.org

Contribution List

- **Thien Nguyen** — Project lead, data pipeline , feature engineering (`feature_engineering.py`), SHAP analysis, literature review (LightGBM, SHAP sections), report writing and coordination.
- **Peter Nguyen** — Model training and evaluation , XGBoost and stacking ensemble implementation, hyperparameter tuning, literature review (XGBoost, Gradient Boosting sections).
- **Henry Bui** — Concept drift simulation (`concept_drift.py`), geospatial and regime visualization, discussion and limitations analysis, literature review (Concept Drift section), presentation.